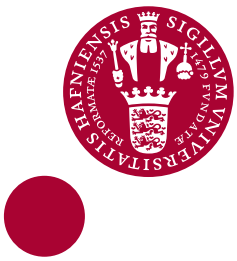


Iterative ratio method

a formal justification for the potentials of mean force

Valentin, Jan; Andreetta, Christian; Paluszewski, Martin; Borg, Mikael; Frellsen, Jes; Paulsen, J.; Boomsma, Wouter Krogh; Bottaro, Sandro; Ferkinghoff-Borg, Jesper; Hamelryck, Thomas Wim

Publication date:
2011

Document version
Peer reviewed version

Iterative Ratio Method

a formal justification for the potentials of mean force

J. B. Valentin*, C. Andreetta*, M. Paluszewski*, M. Borg*, J. Frellsen*, J. Paulsen*, W. Boomsma^, S. Bottaro^, J. Ferkinghoff-Borg^, T. Hamelryck*

* Bioinformatics Center, Department of Biology, University of Copenhagen, Copenhagen, Denmark
^ Biomedical Engineering, Technical University of Denmark (DTU) Elektro, Technical University of Denmark, Lyngby, Denmark

In silico modeling is a valuable tool to investigate biomolecolar structures, with Monte Carlo Markov Chain simulations conquering a major role in this field. The lack of a rigorous justification for the usage and the formalization of their reference states has been a serious topic of discussion for the past decades.
A rigorous solution is here provided, unifying both physical and statistical interpretations, and extending the formalization from empirical potentials to general functional forms.
For the first time, a clear framework is proposed allowing for definition and determination of the reference state of a set of simulations: statistics from previous runs can be employed to iteratively sculpt an energy funnel around the region of interest. Here we show an application to the well known problem of protein folding and design.

The distribution $Q(X)$ over a fine grained variable X, is combined with a probability distribution $P(Y)$ over a coarse grained variable $Y = f(X)$ [1].

$Q(X)$ could represent the information embodied in a fragment library ($\mathcal{F}$), a model of local structure ($\mathcal{T}$) or an energy function ($\mathcal{E}$); Y could be a global descriptor like the radius of gyration, the hydrogen bond network, or the set of pairwise distances.

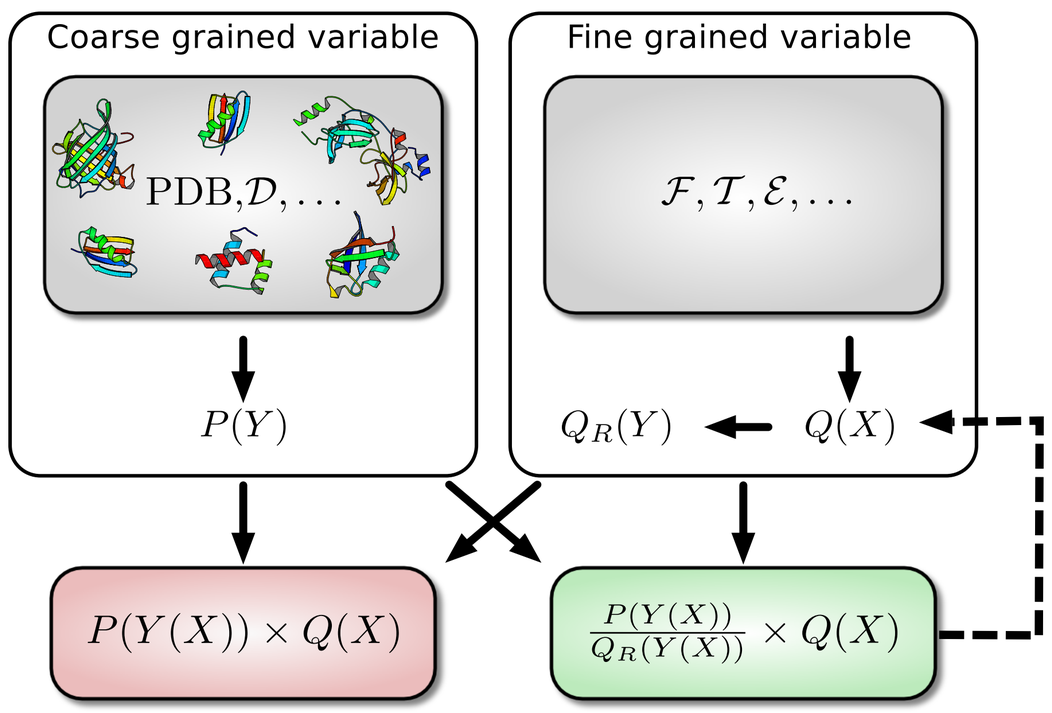
$P(Y)$ reflects the distribution of Y in a database of known protein structures (PDB) or experimental data ($\mathcal{D}$).

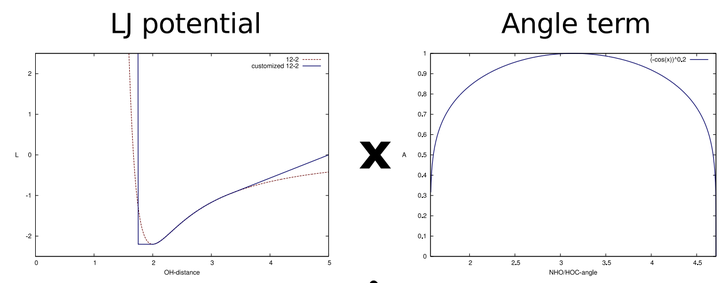
Direct sampling from $Q(X)$ results in a distribution $Q_R(Y)$ that differs from $P(Y)$. The *reference* distribution $Q_R(Y)$ in the denominator (green box) corresponds to the contribution of the reference state in a classical potential of mean force.

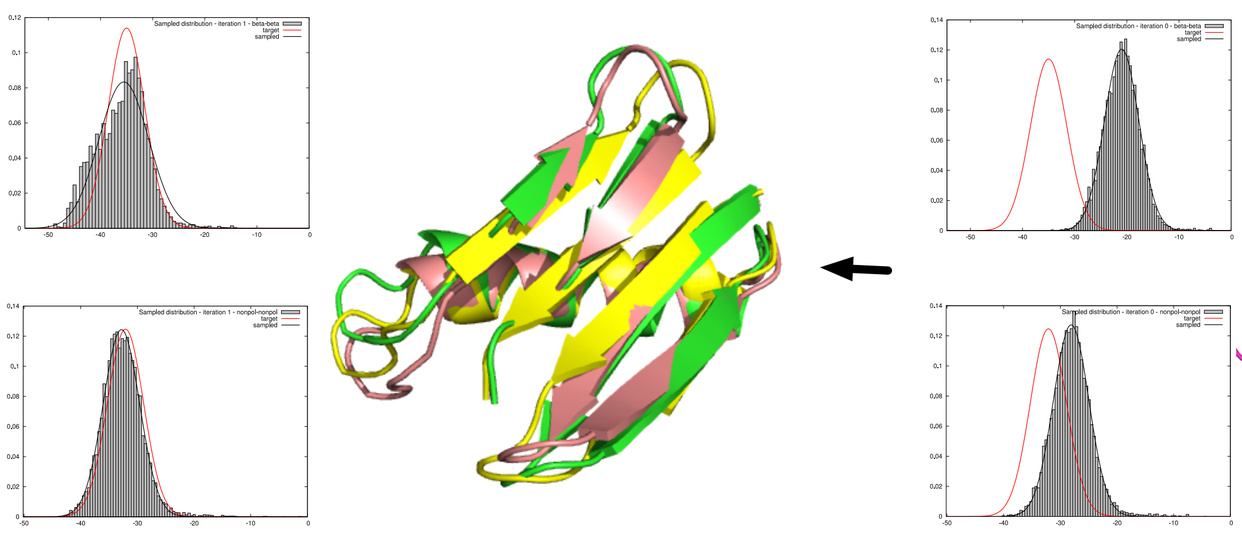
Provided $Q(X)$ and $P(Y)$ are valid probability distributions, the method can be applied iteratively to refine the intial estimate for $Q_R(Y)$, which in the first step can also be assumed uniform.

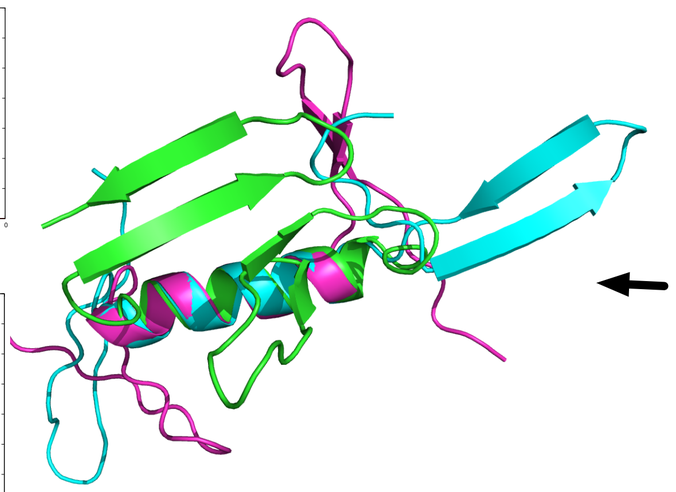
Below we show an example of an iteratively sculpted funnel using a global descriptor Y derived by a customized Lennard-Jones potential [5]. Hydrophobic and hydrophilic contributions are modelled with Normal distributions.

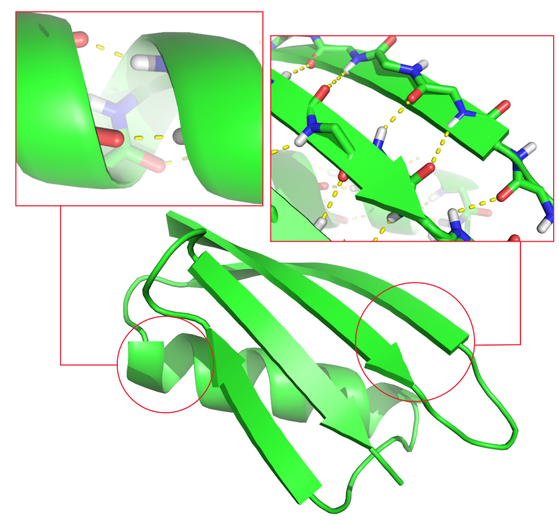
The local structure ($\mathcal{T}$) is proposed by the TorusDBN model [2], trained with the Mocapy++ software package [3]. The reference state can be efficiently estimated from samples obtained from previous iterations, by use of generalized ensemble Monte Carlo methods [4].











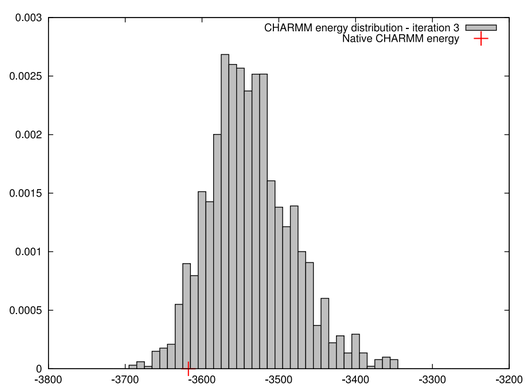
The proposed coarse grained energy function Y categorizes the hydrogen bond network (HB) in different classes C, and assigns a value Y_C to each of them. A total of nine categories are accounted: five cover the backbone bonds, such as β–β and α–α bonds, three categories account sidechain-sidechain interactions and one category is assigned to bonds between sidechain and backbone. We model the distributions $P(Y_C)$ as independent Normals. No other global descriptors (such as secondary structure constraints) are assigned.
Y_C is a sum of modified Lennard-Jones potentials L, multiplied by an angle term A:

$$Y_C(X) = \sum_{j \in HB_C} L_j(X) A_j(X),$$

where j runs over hydrogen bonds in C. The iterative ratio has the following form:

$$Q_{i+1}(X) = \frac{P(Y(X))}{Q_{R,i}(Y(X))} Q_i(X),$$

where $Q_0(X) = Q(X)$ and the initial reference state $Q_{R,0}(Y(X))$ is uniform. This simple set of descriptors already allows to reach native-like energetic states: here we show the output of the popular united-atoms CHARMM19 force field [6].



[1] Hamelryck, Borg, Paluszewski, Paulsen, Frellsen, Andreetta, Boomsma, Bottaro, Ferkinghoff-Borg. *Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized*. 2010. PLoS ONE 5 (11): e13714
[2] Boomsma, Mardia, Taylor, Ferkinghoff-Borg, Krogh, Hamelryck. *A generative, probabilistic model of local protein structure*. 2008. Proc. Natl. Acad. Sci. USA, 105, 8932-8937
[3] Paluszewski, Hamelryck. *Mocapy++ - A toolkit for inference and learning in dynamic Bayesian networks*. 2010. BMC Bioinformatics, 11:126
[4] Frellsen, Ferkinghoff-Borg. *Muninn: An automated method for Monte Carlo simulations in generalized ensembles*. To be submitted
[5] Kabsch, Sander. *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. 1983. Biopolymers 22 (12): 2577–637
[6] Brooks, Bruccoleri, Olafson, States, Swaminathan, Karplus. *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations*. 1983. J Comp Chem 4 (2): 187–217